Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin

Roger A. Hoskins[1*], Joseph W. Carlson[1*], Cameron Kennedy[1], David Acevedo[1], Martha Evans-Holm[1], Erwin Frise[1], Kenneth H.Wan[1], Soo Park[1], Maria Mendez-Lago[2], Fabrizio Rossi[3], Alfredo Villasante[2], Patrizio Dimitri[3], Gary H. Karpen[1,4], and Susan E. Celniker[1]

*These authors contributed equally to this work.

[1]Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, MS 64R0121, 1 Cyclotron Road, Berkeley, CA, 94720 United States

[2]Centro de Biologia Molecular Severo Ochoa,CSIC-UAM, Cantoblanco, 28049, Madrid, Spain.

[3]Dipartimento di Genetica e Biologia Molecolare 'Charles Darwin', Universita 'La Sapienza', Piazzale Aldo Moro 5, 00185 Roma, Italy

[4]Department of Molecular and Cell Biology, University of California, Berkeley, CA, 94720 United States

Running title: *D. melanogaster* Heterochromatin Sequence

Keywords: heterochromatin, *Drosophila*, repeated DNA, sequence assembly, BAC, FISH

Abbreviations: bacterial artificial chromosome (BAC), base pair (bp), Drosophila Heterochromatin Genome Project (DHGP), fluorescence *in situ* hybridization (FISH), kilobase (kb), Megabase (Mb), sequence tagged site (STS), transposable element (TE), whole genome shotgun (WGS)

Address for Correspondence: Lawrence Berkeley National Lab. 1 Cyclotron Road, MS 84-171 Berkeley, CA 94720. 510.486.5034 (office), 510.486.4229 (fax), karpen@fruitfly.org

Genome sequences for most metazoans are incomplete due to the presence of repeated DNA in the pericentromeric heterochromatin. The heterochromatic regions of *D. melanogaster* contain 20 Mb of sequence amenable to mapping, sequence assembly and finishing. Here we describe the generation of 15 Mb of finished or improved heterochromatic sequence using available clone resources and assembly and mapping methods. We also constructed a BAC-based physical map that spans approximately 13 Mb of the pericentromeric heterochromatin, and a cytogenetic map that positions approximately 11 Mb of BAC contigs and sequence scaffolds in specific chromosomal locations. The integrated sequence assembly and maps greatly improve our understanding of the structure and composition of this poorly understood fraction of a metazoan genome and provide a framework for functional analyses.

Heterochromatin is a major component of metazoan and plant genomes (e.g. ~20% of the human genome) that regulates chromosome segregation, nuclear organization and gene expression (*1-4*). A thorough description of the sequence and organization of heterochromatin is essential for understanding the essential functions encoded within this enigmatic region of the genome. However, difficulties in cloning, mapping, and assembling regions rich in repetitive elements has hindered detailed genomic analysis of heterochromatin (*5-7*).

The fruit fly *Drosophila melanogaster* is a major model organism for heterochromatin studies. Approximately 1/3 of the genome is considered heterochromatic, and is concentrated in the pericentromeric and telomeric regions of the chromosomes (X, 2, 3, 4 and Y) (*5, 8*). The haploid female genome contains ~50 Mb of heterochromatin, and males contain an additional ~40 Mb in the entirely heterochromatic Y chromosome (*8*). The approximate cytogenetic boundary between the centric heterochromatin and the euchromatin on each chromosome arm was defined by fluorescent in-situ hybridization (FISH) localization of BACs to mitotic chromosomes (*6*). The heterochromatin is composed of tandemly repeated simple sequences (including satellite DNAs) (*9*), middle repetitive elements (such as transposable elements (TEs) and ribosomal DNA), and some single-copy DNA (*10*)).

The starting material for the Drosophila Release 5 heterochromatin sequence assembly was Release 3 Whole Genome Shotgun draft sequence (WGS3) (*5, 6*). We undertook a retrospective analysis of these WGS3 scaffolds (Supplemental Data 1). WGS3 is an excellent assembly of the Drosophila euchromatic sequence, but has lower contiguity and quality in the repeat-rich heterochromatin. Moderately repetitive sequence, such as transposable elements, are well represented in WGS clones and sequence reads, but tend to be assembled into shorter scaffolds with many gaps and low quality regions. The typical WGS heterochromatic scaffold is significantly smaller than a typical WGS euchromatic scaffold. The N50 length for scaffolds not connected to a chromosome arm ranged from 4 kb to 35 kb, compared to 13.9 Mb for WGS scaffolds that mapped to euchromatic arms (*5*). The WGS3 heterochromatic scaffolds have 5.8-fold more sequence gaps per Mb than the euchromatic scaffolds, as well as lower overall quality. The WGS heterochromatic scaffolds contain a high density of TEs and TE fragments, which typically resulted in sequence gaps due to the difficulty of accurately assigning data to a specific copy of a repeat.

The WGS3 sequence provided the foundation for finishing and mapping heterochromatic sequences and elucidating the organization and composition of the non-satellite DNA in Drosophila heterochromatin. Improved sequence assembly and finishing was dependent on the availability of 10-kb genomic clones and paired end reads (mate pairs) used

in the WGS3 assembly (provided by Celera Genomics), representing 15X clone coverage of the genome (Supplemental Methods). Sequence finishing of small gaps and low quality regions was performed using 10-kb plasmid templates. Higher-level sequence assembly into Mb-sized linked scaffolds utilized relationships determined from BAC-based sequence tag site (STS) physical mapping (see below) and BAC end sequences. In addition to the WGS data, we incorporated finished sequence from 15 BACs (3.4 Mb) that were originally sequenced as part of the euchromatin sequencing efforts (*5, 10*).

Sequence assembly and finishing resulted in significantly fewer gaps, longer scaffolds, and higher quality sequence relative to WGS3 (Supplemental Figure 1). Approximately 15 Mb of this sequence has been finished or improved, and 50% of the sequence is now in scaffolds greater than 378 kb (N50). A summary of the Release 5 sequence assembly statistics by chromosome arm is presented in Table 1. Improved sequence was generated for 145 WGS3 scaffolds, and a set of 90 new scaffolds were produced by joining or filling 694 gaps of previously unknown size between WGS3 scaffolds. The relationships between the initial WGS scaffolds and the improved Release 5 scaffolds can be complex (Figure 1 and Supplemental Figures 2-7); for example, there were eight cases of small scaffolds used to fill gaps within larger scaffolds, and two scaffolds whose gaps interdigitated. As expected, the sequence consists largely of nests of fragmented TEs, and most remaining gaps are bounded by TEs or simple sequence repeats, including simple repeats not previously described (Figure 2). The quality of the improved sequence was measured by calculating the estimated error rates within 10 kb sliding windows (overlapping by 5 kb) [and 1 kb sliding windows (overlapping by 500 bp)] on the consensus sequences (Supplemental Methods). For all but 11 of 1,832 10kb regions not overlapping one of the known TEs, the estimated error rate is less than 1 per 17,986 bp, below the accepted standard for finished genomic sequence of 1 error per 10,000 bp. Continuation of the sequencing efforts will improve quality for these 11 regions.

Concurrent with the sequence finishing effort, we constructed an integrated physical and cytogenetic map to describe the overall structure of the pericentromeric heterochromatin. This map was essential for ordering, orienting and linking WGS sequence scaffolds into larger BAC contigs and Release 5 scaffolds. Heterochromatic sequences present at the centric ends of the Release 3 arm sequences were represented in BAC-based physical maps of the euchromatic and telomeric portions of the chromosomes (*11, 12*). However, most heterochromatic scaffolds had not been mapped in large-insert clones or localized to specific sites on the chromosomes.

BAC-based STS content mapping of WGS3 scaffolds, using 354 probes designed from genomic sequence and five BAC libraries (Supplemental Methods and Supplemental Data 2) extended and linked many scaffolds into larger BAC contigs. The BAC map incorporates scaffolds spanning 13.4 Mb of the WGS3 assembly and links 14 WGS3 scaffolds to the Release 3 arm sequences (Table 2). In regions proximal to the arm assemblies, it links 130 WGS3 scaffolds into 25 multi-scaffold BAC contigs and produced 21 single-scaffold BAC contigs (Table 2, Supplemental Data 3). The largest BAC contig links 20 WGS3 scaffolds spanning 1.7 Mb.

We mapped BAC contigs and sequence scaffolds to specific cytogenetic locations in mitotic chromosomes using FISH (*13*) (Supplemental Methods and Supplemental Data 4). The high repeat content of heterochromatin required the use of single-copy probes (*P*-element insertions from our KV collection (*14, 15*) and cDNA clones from the Drosophila Gene Collection (*16, 17*)) that could be assigned to specific sequence scaffolds. We also used BAC probes that had sufficient single copy sequences to provide unambiguous localizations (Supplemental Data 4 and Supplemental Figure 8). The physical and cytogenetic mapping

results and previously published data were used to produce an integrated map of pericentromeric heterochromatin (Supplemental Data 3). We present cytogenetic locations for 15 BAC contigs linking 80 scaffolds that span 11.2 Mb of pericentromeric heterochromatin in the WGS3 assembly, in addition to the 14 scaffolds that were linked to chromosome arms (Table 2). Currently unlocalized are 50 WGS3 scaffolds in 31 BAC contigs, and an additional 63 WGS3 scaffolds larger than 15 kb that are not represented in the BAC map. Five scaffolds larger than 15 kb and not represented in the BAC map were incorporated into Release 5 by sequence finishing (Supplemental Data 3).

Integration of the map and sequence finishing information resulted in three classes of Release 5 heterochromatic scaffolds: 1) contiguous with the assembled euchromatic arms and extending them farther into pericentromeric heterochromatin (*chromosome arm* 'h'), 2) mapped to specific chromosome arms with partial information on order and orientation and concatenated into "arm" files (*chromosome arm* 'Het'), and 3) unmapped and concatenated into a single file (arm 'U'). The improved, mapped Release 5 scaffolds are diagrammed relative to the chromosome arms in Figure 3, and analysis of sequences and maps by chromosome are in Supplemental Data 5 and 6.

Current technology is capable of determining the genomic sequence, physical associations, and cytological locations of clonable regions of genomes, in this case approximately one-third of Drosophila heterochromatin that precludes the satellite DNA repeats. We have demonstrated significant progress towards our goal of assembling and mapping the components of heterochromatin that are not simple repeats, and show that heterochromatic regions containing single copy genes and a high density of transposable elements can be assembled into high quality, contiguous sequence. The integration of heterochromatic sequence and map data in Release 5 serves as a foundation for more detailed analysis of the structure, function, and evolution of heterochromatic genes and repeats, as demonstrated by the annotations reported in Smith et al. (*18*), as well as analysis by other groups (*19*).

How can we generate an even more complete genomic understanding of Drosophila heterochromatin? The tiling path of overlapping BACs spanning the Release 5 sequence (Supplemental Data 3) provides templates for gap closure and scaffold extension in the regions that contain middle-repetitive elements and single-copy genes. More progress can also be made in localize currently unmapped sequences by performing FISH with additional cDNAs, BACs, and transposon insertions from other collections (*20, 21*). Restriction fingerprints of tiling path BACs will also provide an independent benchmark to evaluate the accuracy of finished sequence assemblies (*22*). However, as noted in Figure 2, several sequence gaps between contigs are bounded by simple satellite DNA repeats. The apparent absence of BACs covering various remaining gaps likely reflects the presence of extensive simple sequence arrays, which are unlikely to be completely closed as the map and sequence are improved. New technologies will be required to determine the sequence and structure of these highly repetitive regions. Our ultimate goal is to produce a complete map and sequence assemblies of the single-copy and middle-repetitive components of the heterochromatin, combined with cytological definition of the locations and structures of large blocks of tandemly repeated, simple sequence DNA.

Our results suggest that elucidating the organization and composition of heterochromatic regions in other organisms is an attainable goal. However, our ability to significantly improve the sequence and maps required three critical components: (1) a high quality WGS sequence assembly, (2) a high depth collection of precisely sized and aligned genomic clones for sequence finishing and gap closure, and (3) physical and cytogenetic

mapping to deduce relationships between WGS scaffolds. The STS content mapping experiments benefited greatly from the availability of large-insert BAC libraries produced by fragmenting genomic DNA with three different restriction enzymes and with physical shearing. Analysis of heterochromatin in other genomes would also benefit from improved algorithms that can successfully and accurately assemble sequence of regions rich in repeated DNA. Genomic information about heterochromatin in multiple species would help elucidate the evolution of genome structure, including histories of rearrangements and translocations, and definitions of conserved blocks of synteny that might point to functionally constrained chromatin domains.

Acknowledgements

We thank Antonio Bernardo de Carvalho for helpful discussions of the Y chromosome sequences and Robert Svirskas, Ashley M. Ryles, Eugene Kym, Raymond Chetty and Sam Galle for technical assistance. BAC-based shotgun sequencing was supported by NIH grant P50-HG00750 to G. M. Rubin and Department of Energy contract DE-AC0376SF00098 to SEC. The majority of this work was supported by NIH grant R01 HG00747 to GHK.

References

1. A. F. Dernburg *et al.*, *Cell* **85**, 745 (May 31, 1996).
2. G. H. Karpen, M. H. Le, H. Le, *Science* **273**, 118 (Jul 5, 1996).
3. P. B. Talbert, S. Henikoff, *Nat Rev Genet* **7**, 793 (Oct, 2006).
4. L. L. Wallrath, *Curr Opin Genet Dev* **8**, 147 (Apr, 1998).
5. S. E. Celniker *et al.*, *Genome Biol* **3**, RESEARCH0079 (2002).
6. R. A. Hoskins *et al.*, *Genome Biol* **3**, RESEARCH0085 (2002).
7. E. E. Eichler, R. A. Clark, X. She, *Nat Rev Genet* **5**, 345 (May, 2004).
8. B. John, G. L. G. Miklos, *The eukaryote genome in development and evolution.* (Allen & Unwin, London, 1988), pp. 416.
9. A. R. Lohe, D. L. Brutlag, *Proc Natl Acad Sci U S A* **83**, 696 (Feb, 1986).
10. M. D. Adams *et al.*, *Science* **287**, 2185 (Mar 24, 2000).
11. R. A. Hoskins *et al.*, *Science* **287**, 2271 (Mar 24, 2000).
12. J. P. Abad *et al.*, *Mol Biol Evol* **21**, 1613 (Sep, 2004).
13. M. Gatti, S. Pimpinelli, *Annu Rev Genet* **26**, 239 (1992).
14. A. Y. Konev *et al.*, *Genetics* **165**, 2039 (Dec, 2003).
15. C. M. Yan, K. W. Dobie, H. D. Le, A. Y. Konev, G. H. Karpen, *Genetics* **161**, 217 (May, 2002).
16. Drosophila Gene Collection.(http://www.fruitfly.org/EST/index.shtml, 2007).
17. M. Stapleton *et al.*, *Genome Res* **12**, 1294 (Aug, 2002).
18. C. D. Smith, S. Shu, C. J. Mungall, G. H. Karpen, *Science* (submitted).
19. J. Brennecke *et al.*, *Cell* **128**, 1089 (Mar 23, 2007).
20. H. J. Bellen *et al.*, *Genetics* **167**, 761 (Jun, 2004).
21. S. T. Thibault *et al.*, *Nat Genet* **36**, 283 (Mar, 2004).
22. M. A. Marra *et al.*, *Genome Research* **7**, 1072 (1997).
23. M. Gatti, S. Bonaccorsi, S. Pimpinelli, *Methods Cell Biol* **44**, 371 (1994).
24. J. Carlson *et al.* (http://www.bdgp.org/sequence/release5genomic.shtml, 2006).

Figure Legends

Figure 1. Comparison of WGS sequence scaffolds to the corresponding Release 5 sequence scaffold. The WGS scaffolds (grey, same orientation; tan, opposite orientation) are diagrammed above the Release 5 scaffold (blue). The thin horizontal lines in the WGS scaffolds represent sequence gaps, and spaces represent clone gaps. Comparisons for all scaffolds are in Supplemental Figures 2-7.

Figure 2. The sequenced regions of *D. melanogaster* pericentromeric heterochromatin. The heterochromatin extends proximally from the euchromatin (black) and includes sequenced and assembled regions (aqua) and unsequenced regions (gray). The actual gap sizes between sequence scaffolds are unknown and are presented with an arbitrary 0.5 Mb separation. Finished or improved scaffolds, which end in known or novel simple repeats, are shown with the terminal repeat sequence indicated. The scaffold CP000217, originally identified as part of 2RHet but subsequently mapped to 3LHet, is shown here at its updated location (see text).

Figure 3. Integrated map of *D. melanogaster* pericentromeric heterochromatin. The cytogenetic reference map of the heterochromatic regions of the chromosomes with numbered divisions (h1-h61) and centromeres (C) is shown (*23*). Release 5 sequence scaffolds (green) are indicated at their cytogenetic map locations and labeled with their GenBank accession numbers. Scaffolds (11 Mb in total; see scale bar) and the heterochromatin (90 Mb in total) are represented at different scales. Sequence contigs (thick bars) and sequence gaps (thin bars) within scaffolds are shown. Some sequence gaps are too small to be represented at this scale. A clone gap in the 2Lh sequence is indicated (thin red bar). Joins between Release 5 scaffolds present in the BAC map assembly but not yet incorporated in the sequence assembly are shown (thin blue bars). Cytogenetic locations are indicated by lines (gray) connecting scaffolds to cytogenetic ranges. The heterochromatin-euchromatin boundaries within the sequence of the chromosome arms, based on BAC FISH (*6*), are indicated (dashed fuscia lines). The orientations of Het scaffolds are not necessarily known (Supplemental Data 6; (*24*). CP000217, originally identified as part of 2RHet but subsequently mapped to 3LHet, is shown here at its updated location; CP000206, originally identified as part of 3RHet but subsequently removed to the unlocalized scaffolds, is not shown (Supplemental Data 1).

Table 1: Status of Release 5

Region	Size (bp)	BAC-Based Rel. 5	Rel. 5 w/o N's	N50	Sized Gaps	Total Gap Size	Unsized Gaps
Xh	392,502	312,439	392,502	392,502	0	0	0
XHet	204,112		204,112	204,112	0	0	0
2Lh	1,010,570	1,010,470	1,010,470	591,203	0	0	1
2LHet	368,872		297,872	99,162	2	71,000	0
2Rh	1,285,689	973,874	1,285,689	1,285,689	0	0	0
2RHet*	3,288,761		2,721,941	244,298	17	566,020	8
3Lh	1,587,982	1,020,114	1,587,982	1,587,982	0	0	0
3LHet*	2,555,491		2,416,308	366,456	12	138,483	7
3Rh	378,656	378,656	378,656	378,656	0	0	0
3RHet	2,517,507		2,264,306	252,624	10	252,801	4
YHet	347,038		242,806	9,129	30	101,632	26
Unmapped modified	2,419,890		2,222,443	73,591	15	194,247	32
Total for modified sequence	16,357,070		15,025,087	378,616	86	1,324,183	78
Unmapped unmodified	7,629,047		6,145,805	2521	439	1,239,942	2433
Total	**23,986,117**	**3,383,114**	**21,170,892**		**525**	**2,564,125**	**2511**

The sequence statistics for the chromosomes is divided into regions contiguous with the euchromatic arm sequences (e.g. Xh) and the regions mapped cytologically to those chromosome arms but not presently connected to the rest of the sequence of the arm (e.g. XHet). Bac-Based Rel. 5 refers to the amount of heterochromatin finished in BACs. *Statistics for 2RHet and 3LHet reflect the sequence distributed as Release 5 of the genome and does not account for the scaffold CP000217 moved from 2RHet to 3LHet subsequent to the release.

Table 2. Summary of the integrated physical and cytogenetic map assembly

Chromosome Arm	STSs in BAC Map	Failed STSs in Mapped Contigs	WGS3 Scaffolds Linked to Chr. Arm	WGS3 Scaffolds in Het Contigs	Het Contigs	Sum of WGS3 Lengths in Mapped Scaffolds (kb)
XL	16	0	3	2	1	498
2L	28	1	2	5	1	1,018
2R	91	1	5	29	3	3,517
3L	91	1	3	24	6	4,039
3R	51	0	0	20	4	2,101
4R	8	2	1	0	0	65
Y	1	4	N/A	0	0	0
Subtotal (localized)	286	9	14	80	15	11,238
U	68	N/A	N/A	50	31	2,177
Total	**354**	**9**	**14**	**130**	**46**	**13,415**

Supplemental Figures and Data

Supplemental Figure 1. Assembly of a challenging region of 3RHet

Supplemental Figures 2-7. Comparison of WGS sequence scaffolds to the corresponding Release 5 sequence scaffolds

Supplemental Figure 8. Cytogenetic localization of BAC clones

Supplemental Data 1. Description of the WGS heterochromatin sequence

Supplemental Data 2. STS probes and hybridization summary

Supplemental Data 3. Integrated physical and cytogenetic map of *D. melanogaster* pericentromeric heterochromatin

Supplemental Data 4. FISH localization of P-element insertions, cDNAs and BACs

Supplemental Data 5. Analysis of Release 5 heterochromatic sequence by chromosome

Supplemental Data 6. Analysis of the integrated physical and cytogenetic map of pericentromeric heterochromatic by chromosome arm